Traffic Measurement Study on Video Streaming with the Amazon Echo Show

Frank Loh, Viktoria Vomhoff, Florian Wamser, Florian Metzger, Tobias Hoßfeld
firstname.lastname@informatik.uni-wuerzburg.de
University of Würzburg, Institute of Computer Science, Germany

ABSTRACT

The Amazon Echo Show is one of the most widely used smart speakers with the ability to stream video. Due to its popularity, the traffic profiles of such devices are of interest to network operators and providers. This work presents a measurement study of the Amazon Echo Show in terms of network traffic and streaming behavior. More than 470 hours of streaming data are collected and analyzed at network layer. Based on this, streaming quality is derived at application layer. The study quantifies the traffic and shows that streaming with the Amazon Echo Show is comparable to streaming with a native web browser, but in a more conservative way.

ACM Reference Format:
Frank Loh, Viktoria Vomhoff, Florian Wamser, Florian Metzger, Tobias Hoßfeld. 2019. Traffic Measurement Study on Video Streaming with the Amazon Echo Show. In *4th Internet-QoE Workshop: QoE-based Analysis and Management of Data Communication Networks (Internet-QoE'19), October 21, 2019, Los Cabos, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3349611.3355543

1 INTRODUCTION

Smart speakers are currently the fastest growing segment of consumer technology in the world [3]. In 2017, the market grew from 2.9 million to 9 million devices. With the release of the Amazon Echo Show, the feature set of smart speakers has evolved from simple voice-controlled Internet queries and home device controlling to video streaming.

There are already many studies analyzing video streaming performance [9, 12] or providing large traffic datasets for complex analysis [7]. For smart speakers, however, a deep understanding of the behavior is missing. With the increasing amount of devices, details about the generated network traffic and the streaming behavior are of high interest.

From a network operator's point of view, it is essential to quantify the generated traffic for different streaming situations, platforms, and user behavior with smart speakers. Thereof, streaming quality for an end user can be analyzed by mapping traffic characteristics to application layer behavior. Video startup delays, interruptions, and quality degradations can be detected and appropriate mechanisms may be implemented within the network or at the application layer. For that reason, a profound understanding of the streaming behavior and the resulting network traffic is important.

This work presents a traffic measurement study on video streaming with the Amazon Echo Show. First, the ability to understand voice commands is analyzed over time. Then, the generated network traffic when requesting videos in different bandwidth situations is analyzed. A study is conducted on more than 470 hours of captured streaming data. The streaming behavior is determined and mapped to application layer parameters like playback quality or downloaded video seconds per request. These results are compared to streaming with a native web browser and a smartphone browser.

The contribution of this work includes a setup for video measurements with the Amazon Echo Show. The network traffic is captured and postprocessed to illustrate the comprehension of voice commands over time. Furthermore, streaming parameters with the device are analyzed based on the network traffic. Finally, conclusions are drawn to application layer parameters which may be the baseline for subjective user studies to analyze Quality of Experience. The captured network traffic and streaming behavior can be used by network operators to extend their traffic models.

The remainder of this work is as follows: in Section 2 background information and related work is summarized. Section 3 presents information about the testbed and the conducted study. Discussion of the measurement results follow in Section 4. Finally, in Section 5 conclusions are drawn.

2 BACKGROUND

This section summarizes fundamental background required to understand this work. First, the Amazon Alexa ecosystem is introduced followed by the description of the Echo devices.

Internet-QoE'19, October 21, 2019, Los Cabos, Mexico

ACM ISBN 978-1-4503-6927-5/19/10…$15.00
https://doi.org/10.1145/3349611.3355543

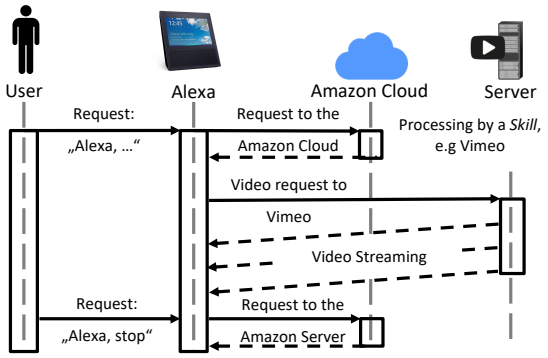

Figure 1: Alexa usage and connection establishment

2.1 Amazon Alexa Ecosystem

Amazon's Alexa devices were the first dedicated home devices to support conversational user interfaces and cloud based artificial intelligence that reached the consumer market. All devices have at least one speaker with microphone that can be used among others for voice commands, controlling smart home devices or stream music. A touch display is integrated in the Echo Spot and Show where videos can be played by Amazon Video or other streaming services. Furthermore, smart speakers from Amazon benefit from the data and device ecosystem that learns from different devices.

2.2 Usage and Communication

In this work, we measure and examine video playback based on network layer investigations for the Amazon Echo Show. Currently Prime Video, Vimeo, and Dailymotion is supported as streaming backend with a specific app. It is possible to request video content with a voice command recorded by eight microphones of the device. The requested videos are displayed on the 7-inch, 1024 x 600 pixels screen. Figure 1 shows an exemplary video playback as a sequence diagram. First, the user requests a video. The Echo sends the request as an audio stream to the Amazon Cloud while the voice command is being spoken. There, it is processed from voice to text. The request is converted into a video request and sent to the specific platform the video is available at. As soon as Alexa receives a reply, it connects to the respective video server. When the user prompts Alexa to stop playing video, the voice command is submitted to the Amazon Cloud and processed. If successful, Alexa then returns to the home screen.

2.3 Related Work

Research about smart speakers to understand and fulfill a wide range of tasks is often directly translated to user retention [2] or satisfaction [5]. A focus in [10] is user satisfaction of the Amazon Echo devices. Since many voice assistants featured on smart speakers are installed on different types of devices, research exists that investigates the general capabilities of smart assistants [6]. Furthermore, the authors of [1] investigate if voice assistants understand people with speech impairments, in this case Dysarthria, and come to mixed results. Smart speakers are placed in privacy sensitive places in one's own home, making security analyses considerations critical. In [8] a study of the privacy concerns and resulting usage behavior of smart speaker users is done. Actual security analyses and attacks were performed, e.g., in [4]. Because this work is examining the Echo Show in particular, it is also worthwhile to consider existing work on adaptive video streaming. But while the device is similar in its internal makeup and available resources to common smartphones, the video streaming behavior might still be vastly different since these devices have different objectives, environments, and application ecosystems. To the best of our knowledge no work has been published that analyzes the video streaming behavior of smart speakers. However, HTTP adaptive streaming itself has been well investigated in the past. For example, the characteristics of video streaming traffic are discussed in [11].

3 MEASUREMENT PROCEDURE

In this section, the measurement methodology, investigated scenarios and data postprocessing is described.

3.1 Testbed

To capture the traffic of the Echo Show, an automated measurement setup is created. The Echo is connected to a computer, used as controlling instance, WiFi access point, and for data collection. A speaker is connected to the computer to play pre-recorded voice requests to the Echo Show. The setup is created in an isolated environment to minimize background noise and interference with the measurement that may lead to misunderstanding the spoken voice commands.

For data collection, the testbed is established as follows: the computer, that is connected to the Internet, opens a WiFi access point for the Amazon Echo Show. At the Echo device, the video streaming skills Vimeo and Dailymotion are activated. The skills are freely available and no Amazon Video account is required. For each measurement, a pre-recorded audio file is played out to request content by the Amazon Echo Show. Requesting a video is done with the command: "Alexa, play video *videoname* from *streaming platform*". The complete traffic is captured with *tcpdump* at the computer. By limiting the available network bandwidth with the traffic configuration application *tc*, the behavior of the Echo device under various network conditions is monitored. Before the investigation, the Amazon Echo Show was never used to monitor the voice command comprehension over time.

Table 1: Overview of all studied videos

Platform	Video title	Video ID	Max. avg. bitrate	Available qualities
VI	Big Buck Bunny	1084537	2637.14 kbit/s	240p, 360p, 720p
VI	Kometen - Lichtgeschwindigkeit	90018009	708.74 kbit/s	360p, 720p
VI	Insight: Brad Bird on Animation	189791698	1677.49 kbit/s	360p, 540p, 720p, 1080p
VI	The City Limits	23237102	860.68 kbit/s	240p, 360p, 1080p
VI	Landscapes: Volume Two	29950141	657.57 kbit/s	240p, 360p, 720p, 1080p
VI	Linkin Park: Castle of Glass	58669810	655.42 kbit/s	360p, 1080p
DM	David Coverdale: Love is Blind	x2mdjmz	459.62 kbit/s	144p, 240p, 380p
DM	Metallica: Nothing Else Matters	x13pfm0	462.07 kbit/s	240p, 380p
DM	Bad Bunny feat. Drake: Mia	x6vb8ja	2128.70 kbit/s	144p, 240p, 380p, 480p, 720p, 1080p
DM	The House with a Clock in its Walls	x6gxjpe	2119.73 kbit/s	144p, 240p, 380p, 480p, 720p, 1080p

3.2 Scenario Description

The measurement study with the Echo Show includes video streaming investigations with 4962 individual runs and more than 470 h of video, summarized in Table 1. The playback quality is chosen automatically by the device. The bandwidth settings are as follows: the 400 Mbit/s limitation is selected to show the best case streaming performance of the device not influenced by any bandwidth limitation. According to a previous study of another video platform [7], 3 Mbit/s was shown to be a good limit for playing "smooth" HD video quality, also shown in the *max. avg. bitrate* column in Table 1. There, the average bitrate of the highest played out quality by the device is listed accordingly to the manifest file. According to a previous study 1 Mbit/s shows frequent changes in video quality and occasional video rebuffering events and 800 kbit/s is selected to analyze smaller qualities. The video selection includes six videos from Vimeo, tagged as *VI* in the *Platform* column and four videos from Dailymotion (*DM*). The video duration is between 2:09 min and 9:57 min per video to cover a wide area of typical video clip lengths. Each video was played out completely. Additionally, a mixture of videos with many views and unknown videos is chosen. The goal is to analyze the behavior of the device for different bandwidth limitations and videos from both platforms.

3.3 Data Postprocessing

The data postprocessing is threefold: first, it is checked whether the Echo Show plays out the requested video by analyzing the received traffic and comparing it to the required amount of data and errors in the playback behavior are logged. Second, the network dumps are validated by checking, among others, the timestamp and total size of the dump to detect errors in the monitoring process. Last, network layer data are processed and mapped to application layer metrics based on the manifest file. Details about the postprocessing steps are given in the following.

3.4 Errors over Time

The error occurrence over time is analyzed in 1657 runs in Figure 2 for video 1 in the top plot and video 2 in the bottom one exemplary. The measurements are conducted between May and July 2018. The peaks show error-prone measurements. The x-axis shows all runs normalized, starting with the first run to the left. It is shown that for video 1 most of the 25 % error-prone measurements occurred at the beginning of the capturing period and the errors are detected in bursts. In video 2, 29 % of all runs show errors, while they are more evenly distributed over the whole period. For video 1, more than 50 % of all errors occurred in the first 25 % of all measurement runs, 90 % are detected in the first 70 %. At the end, the voice commands are processed correctly more often and the device switched from idle state to playback correctly. For 70 % of all false measurements, one error is followed by at least a second one, in the first 50 %, it is in 80 %. For video 2, the first 50 % error-prone measurements are conducted in the first 51.57 % and only in 26 % bursts with a length of at least two are detected. This error curve shows an improvement of the requesting behavior. Video 1, "Big Buck Bunny", is misunderstood often at the beginning, playing out the wrong video or no video at all. Thus, the learning process is visible during the measurement for this video. Video 2, in contrast, shows 27.10 % errors. These errors are caused by faults at the end of the measurement or playback errors. There, the video is terminated before the video is ending, although this is not triggered on purpose. For this behavior, no improvement is detected. Thus, although a controlled and deterministic environment is created, differences in measurement errors for different videos and over time are detected. One possible reason is the improvement of content location detection and delivery if it is requested more often. Thus, the amount of error 2 is decreasing. Furthermore, more popular videos are less likely requested wrong. However, this evaluation is very context-specific, only tested with a single speaker and must be analyzed with a larger test set in more detail.

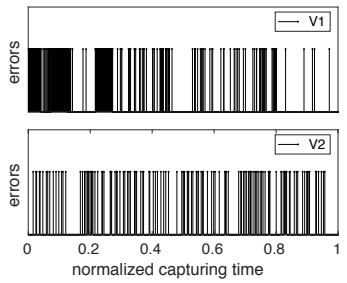

Figure 2: Errors over time (black stems indicate errors)

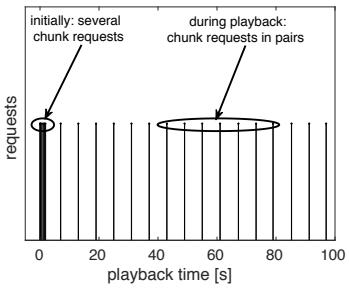

Figure 3: Chunk requests for video 1 for 400 Mbit/s

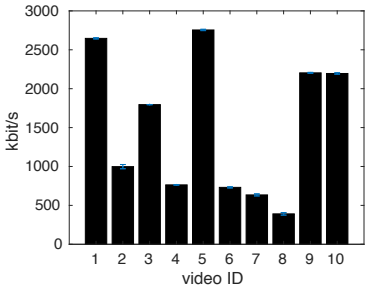

Figure 4: Average downlink per video, streaming with 400 Mbit/s

3.5 Network Layer Data

A connection between the device and an Amazon server is created when requesting data with the Echo Show. The encrypted network layer traffic between both end points passes the measurement computer, where it is captured and analyzed for all open connections. For each packet, the timestamp, the source and destination IP address and network port, the packet size, the network protocol, and additional DNS information are extracted. With encrypted network traffic, no deep packet inspection is possible. Thus, all network flows are determined according to IP address and port number pairs. By looking for the open network flow, where the DNS request is asking for the IP address of *Vimeo.com* or *Dailymotion.com* respectively, the video stream is extracted. For each video stream, the HTTPs protocol is used while only one parallel connection is open. By following this video stream, all uplink requests are logged. In that way, the network layer video chunk requests can be determined, displayed in Figure 3. There, the extracted requests for video 1 streamed with 400 Mbit/s from 0 s to 100 s measurement time are shown exemplary. Initially, a burst of several chunk requests is detected, indicating a video downloading phase to fill the playback buffer. The goal is to fill the buffer until the playback can start as fast as possible to minimize initial delay. Thus, the amount of chunk requests is dependent on the amount of content sent for one request and the initial target buffer level. Afterwards, a repeating pattern of a pair of two chunks each 6 s is detected with two small outlier at 43 s and 61 s. There, the second chunk of the pair is delayed for 0.5 s. If enough bandwidth is available, this behavior is visible for the whole video and all runs to keep the buffer above a certain level. A more detailed description is following in Section 4.

4 STREAMING EVALUATION

The evaluation presented in this section covers video streaming with the Amazon Echo Show based on network layer data and an analysis of resulting application layer properties.

4.1 Traffic Characterization

Figure 4 shows the average downlink rate for all streamed videos with 400 Mbit/s. The required downlink rate is calculated by following the network flow matched to the video stream according to Section 3. The errorbars depict the standard deviation. The figure shows that the maximal required downlink rate is 2700 kbit/s for video 5. This is comparable to the 720p quality video available from Vimeo, according to the manifest file where the average bitrate is listed as 2658 kbit/s. The minimal value is about 700 kbit/s for video 7, and 500 kbit/s for video 8, comparable to the 360p or 380p video quality available on the respective streaming platforms. For example for video 8, it is 462 kbit/s. It is remarkable that, although it is also available in 1080p quality, video 6 is streamed in 360p quality. The same is visible during the study with all other videos. Thus, it is assumed that the Alexa does not request 1080p quality if not triggered manually and only plays the best available quality up to 720p.

Streaming Properties. Figure 5 shows the first 100 s of one measurement run with 400 Mbit/s for video 1 and video 2 to explain the different streaming behavior on network layer. The other runs behaved similarly. Other videos show either the behavior of video 1 or video 2, while both types are detected for Vimeo and Dailymotion. The y-axis shows the normalized cumulated download per video. It is visible that for video 2, the complete video is downloaded within the first 5 s. Video 1 is streamed with periodically requesting new data after a short initial delay phase. Since the same is detected for the other videos, independent on the streaming platform, it is shown that the Amazon Echo Show either downloads the complete video at the beginning of the stream or requests new data when necessary dependent on the video.

To compare this with other players, Figure 6 shows the streaming behavior for video 1 in brown and video 2 in orange conducted in the same network for the Echo Show, the Google Chrome browser installed on a desktop computer and the Google Chrome smartphone browser. For both Chrome

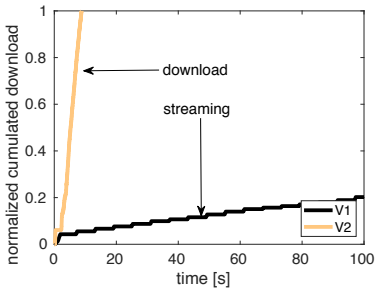

Figure 5: Alexa's streaming behavior for different videos

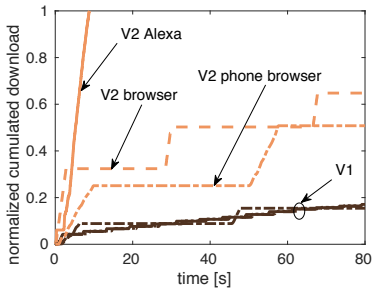

Figure 6: Video streaming behavior for different devices

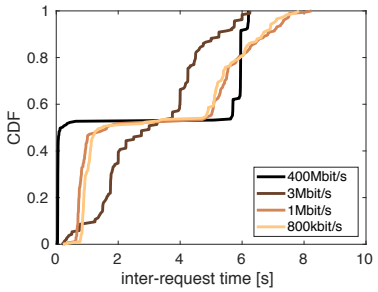

Figure 7: IRT for different bandwidths for video 1

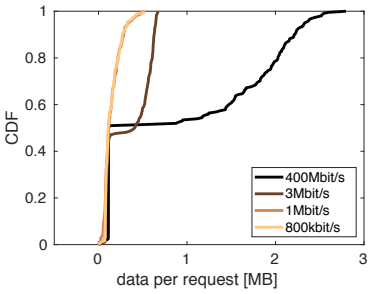

Figure 8: Data per request for different bandwidth limits

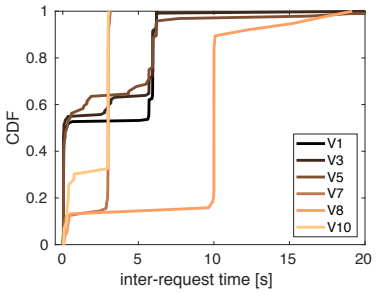

Figure 9: Overview of IRT behavior for different videos

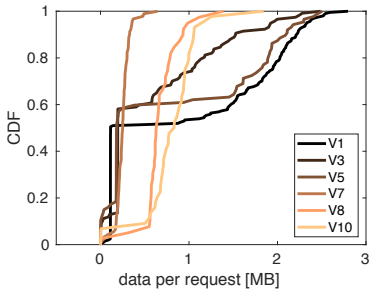

Figure 10: Data per chunk request for different videos

measurements no additional resource intense service is running. The figure shows the normalized cumulated download of the first 80 s of the stream. During the rest of the streaming, the behavior is not changing. Video 1 is streamed with a periodical download of video chunks. The amount of downloaded data per request is similar for streaming with the Alexa or the web browser. Using the smartphone browser, the downloaded video portion is slightly larger. Video 2 is downloaded completely at the beginning by the Alexa. With the Google Chrome browser, large video portions are downloaded but with periods of no download in between. The same behavior is visible with the smartphone browser, but with larger intervals between two consecutive downloading phases. Thus, compared to streaming with the Amazon Echo Show, the video with a web or smartphone browser is streamed with a periodical buffer refill with different sized steps. Thus, streaming with the Echo Show is highly related to the currently played out video.

According to Section 3 chunk requests are sent to the server on network layer. As response, audio and video data are sent to the client. Figure 7 shows the inter-request time (IRT) of chunk requests streaming video 1 with 400 Mbit/s, 3 Mbit/s, 1 Mbit/s, and 800 kbit/s. For 400 Mbit/s, 51 % of all IRTs are smaller than 0.1 s, 49 % are between 5.8 s and 6.2 s. This is an indicator for different request types. Additionally, Figure 6 shows that there are pauses in the download. These pauses are visible before the next chunk is requested, thus when the previous one is completely downloaded. Moreover, dependent on the video quality, audio chunks are much smaller than video chunks, and thus downloaded faster. Assumed that there is no parallel download of multiple chunks of the same type, we conclude that in case of sufficient bandwidth, audio and video chunk requests are sent in pairs. In this way, the buffer at the device can be periodically refilled by the received data. Regarding the other scenarios, this behavior is also visible for the 1 Mbit/s and 800 kbit/s scenario but less obvious. For the 3 Mbit/s scenario, a smaller distinction between different requests is visible observing the IRT. The pairwise requests of audio and video is only possible if enough bandwidth is available. This is the case in the 400 Mbit/s scenario. Based on the IRT, it is not obvious for 3 Mbit/s, 1 Mbit/s, and 800 kbit/s.

The data per request is shown in Figure 8 for video 1 by summing up all data between two chunk requests. For 400 Mbit/s limit, 51 % of all requests contain less than 100 kB. For the other 49 %, between 1 MB and 2.8 MB are downloaded. Regarding the 3 Mbit/s scenario, 45 % of all requests contain less than 100 kB, the other 55 % contain between 0.35 MB and

0.7 MB. Although, no such clear distinction is visible in the IRT distribution, a small request and a larger request class is shown. Since the total amount of data required for the audio stream is smaller for 720p and 360p quality, the small requests are assumed to be audio, while the large ones are video requests. With 400 Mbit/s, more data is requested per chunk. Compared to that, the difference in the 1 Mbit/s and 800 kbit/s scenarios is less distinct. As a consequence, the audio and video stream can be separated next to the IRT by the data between two consecutive requests when the video is streamed with 3 Mbit/s. A different quality must be played out since much larger requests for 400 Mbit/s are downloaded. Comparing the downloaded data with the required data in the manifest file, it is shown that with a bandwidth limitation of 3 Mbit/s the 360p video quality is played out. This is remarkable since the mean bitrate of the 720p quality, for example for video 1, is 2637 kbit/s. Additionally, no adaptation is detected. The 1 Mbit/s and 800 kbit/s scenarios show nearly the same CDF. It is assumed that the same quality is played out. This is also observed, comparing the total amount of downloaded data with the video sizes of different qualities.

Figure 9 shows the IRT distribution for the six videos that are not downloaded completely at the playback start. The differences in the percentage of small and the large IRTs dependent on the video is shown. Video 3 and video 5 are similar to video 1, while about 10 % of all IRTs for both videos are between 1 s and 6 s. For video 7 and video 10, the maximal IRT is 3 s. For video 8, about 10 % of the IRTs are smaller than 1 s and more than 80 % are 10 s or larger. Thus, for all videos a clear distinction between very small and large IRTs is visible while larger IRTs vary between videos. In contrast, Figure 10 shows the amount of data per request as CDF for the same videos. It is shown that for video 1, video 3, and video 5 a clear distinction between small and large requests is visible, that is not that clearly observable for the other videos. Video 8 and video 10 show comparable CDFs, while Figure 9 shows that much more video is downloaded with one request for video 8. Thus, we conclude from network layer data analysis that, although the same bandwidth limit is set, different qualities are streamed, also shown in Figure 4 or by the manifest file.

Summarizing, the request analysis shows that the data per request is different dependent on video and quality. Additionally, the IRTs vary between different videos. It is visible that larger chunks are requested with higher available bandwidth and thus, better video quality. Since large parts of the IRTs are similar while the data per request have a high variance, it is assumed to request new video content based on video seconds to refill the playback buffer and not data in MB. A simplified model looks like this: video is downloaded with probability p. Video is streamed with pairs of periodic requests. The data per request depends on the available bandwidth and the size of the chunks. However, this requires a more detailed analysis of the switching behavior.

5 CONCLUSION

A profound understanding of currently available smart speakers is required since the amount and diversity is steadily increasing. Thus, in this work, network layer traffic measurements are done for a broad range of videos and streaming scenarios. By analyzing the generated traffic, the voice understanding is analyzed and application layer parameters are estimated. Our tests show that the streaming behavior with the Echo Show is very conservative and highly related to the requested video. If the available bandwidth is only a little higher than the average video bitrate, a lower quality level is requested constantly without changing bandwidth, although this would be possible. For future works, a study with variable bandwidth limitations is necessary to completely understand the buffering behavior. This is required to model the complete streaming behavior and thus, the network traffic generation process of the device.

REFERENCES

[1] Ballati, F., Corno, F., and DE Russis, L. "hey siri, do you understand me?": Virtual assistants and dysarthria.

[2] Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. Understanding the long-term use of smart speaker assistants.

[3] Canalys. Google beats amazon to first place in smart speaker market, 2018. Accessed: 2018-11-21.

[4] Chung, H., Park, J., and Lee, S. Digital forensic approaches for amazon alexa ecosystem. *Digital Investigation* (2017).

[5] Hashemi, S. H., Williams, K., El Kholy, A., Zitouni, I., and Crook, P. A. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *27th ACM International Conference on Information and Knowledge Management* (2018).

[6] Hoy, M. B. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical reference services quarterly* (2018).

[7] Karagkioules, T., Tsilimantos, D., Valentin, S., Wamser, F., Zeidler, B., Seufert, M., Loh, F., and Tran-Gia, P. A public dataset for youtube's mobile streaming client. In *TMA Conference* (2018).

[8] Lau, J., Zimmerman, B., and Schaub, F. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.* (2018).

[9] Orsolic, I., Pevec, D., Suznjevic, M., and Skorin-Kapov, L. A machine learning approach to classifying youtube qoe based on encrypted network traffic. *Multimedia tools and applications* (2017).

[10] Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. Alexa is my new bff: social roles, user satisfaction, and personification of the amazon echo. In *CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017).

[11] Rao, A., Legout, A., Lim, Y.-s., Towsley, D., Barakat, C., and Dabbous, W. Network characteristics of video streaming traffic. In *7th Conference on emerging Networking Experiments and Technologies* (2011).

[12] Wamser, F., Casas, P., Seufert, M., Moldovan, C., Tran-Gia, P., and Hossfeld, T. Modeling the youtube stack: From packets to quality of experience. *Computer Networks* (2016).